

Latent variable modelling, assignment 1.

Master DS - AMU

Academic year 2024-2025

Exercise 1: Comparing two proportions

A clinical trial investigated whether an intrauterine device (IUD) emitting progesterone could prevent endometriosis in women with breast cancer undergoing Tamoxifen (TF) treatment. At the end of the study:

- In the IUD group, 5 out of 56 women developed fibroids.
- In the control group, 13 out of 53 women developed fibroids.

Tasks:

- Posterior Distribution for Proportion of Fibroids (PF):
 - Using a uniform prior, calculate the posterior distribution for PF in each group.
 - Compute the 95% credible interval for PF in each group.
- Incorporating Prior Information: A prior clinical study of 20 patients in each group reported: 8 fibroids in the treatment group and 12 fibroids in the control group.

You do not fully trust this prior study but are willing to incorporate it with an equivalent sample size of 10 for each group.

 - Determine the posterior distribution of PF for each group using this adjusted prior.
 - Find the posterior distribution of the difference in PF between the groups.
 - Compute the posterior probability that PF is smaller in the IUD group compared to the control group.
- Normal Approximation:
 - Repeat the above analyses using normal approximations to the posterior distributions.

Exercise 2: normal model with unknown mean and known variance

Load the cavendish dataset in R, which contains 23 experimental measurements of the Earth's density (in g/cm^3) as recorded by the physicist Henry Cavendish. Assume the data follow a normal distribution with an unknown mean, and a known variance of ($\sigma^2 = 0.04$).

Construct a Bayesian model for the unknown mean using a normal prior: $\mathcal{N}(6, \sigma_0^2)$; $\sigma_0^2 = \{0.08, 2\}$.

Tasks:

- Plot the prior and posterior distributions on the same graph.

- Summarize the posterior distribution, including its mean and variance.
- Compute a 95% credible interval for the Earth's density.
- Calculate the posterior probability that the Earth's density is greater than 6.

Exercise 3: identifiability

Let $\theta = (\theta_1, \theta_2)$. Consider a prior $\pi(\theta_1, \theta_2)$. Suppose the likelihood function for the observed data y depends on θ_1 and not θ_2 , i.e., $p(y|\theta_1, \theta_2) = p(y|\theta_1)$.

- find the conditional posterior density of θ_2 given θ_1
- find the marginal posterior density of θ_2
- do you learn anything about θ_2 ?

Exercise 4: pareto

Let $x \sim Pa(\lambda, \theta)$, $\lambda > 0, \theta > 0$, a pareto distribution $p(x|\theta) = \lambda \frac{\theta^\lambda}{x^{\lambda+1}} 1_{\{[\theta, +\infty)\}}(x)$, assume that $\theta \sim Be(\alpha, \beta)$. Show that, if $\lambda < 1$ and $x > 1$, a particular choice of α and β gives that the posterior of θ is uniform on $[0, 1]$.

Exercise 5: about the prior

Given a proper distribution $\pi(\theta)$ and a sampling distribution $p(x|\theta)$. Show that the only case such that the posterior $\pi(\theta|x)$ and $\pi(\theta)$ are identical occurs when $p(x|\theta)$ does not depends on θ .

Exercise 6: HPD regions

Consider a Bayesian model with likelihood $p(y|\theta)$ and prior $\pi(\theta)$ $\theta \in \Theta \subset \mathbb{R}$. Let $\gamma = g(\theta)$ be a one to one transformation of θ . Let A be a $(1 - \alpha)100\%$ HPD region for θ . Define $B := \{\gamma : \gamma = g(\theta), \theta \in A\}$. Is B an $(1 - \alpha)100\%$ HPD region for γ ?

Exercise 7 (*): multivariate normal

The table hereafter gives you the grades of $n = 88$ students in different subjects: mechanics, vectors, algebra, analysis and statistics. You model simultaneously these grades assuming the observations are multivariate normal $N_5(\mu, \Sigma)$.

We define

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$S = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$

First consider that Σ is known.

- Put a multivariate normal prior on μ and compute the posterior.
- Application to the scores data. (Consider the known variance matrix to be the estimated one and choose a prior as you want to.)
 - What is the probability of having a grade lower than the average in statistics while being above average in analytics only ?
 - What is the probability of having a grade lower than the average in statistics while being above average in all the other subjects ?

Additional questions

- Suppose now that μ is known and Σ is unknown. The prior for Σ^{-1} (say of dimension $p \times p$) is the Whishart distribution $W(A, \nu)$, A is a non random $(p \times p)$ positive definite matrix, $\nu > 0$ is a scalar giving the number of degrees of freedom.

$$p(H) = \frac{1}{c_W} |H|^{\frac{\nu-p-1}{2}} |A|^{-\nu/2} \exp \left\{ -\frac{1}{2} \text{tr}(A^{-1} H) \right\},$$

where $c_w = 2^{\frac{\nu p}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^p \Gamma\left(\frac{\nu+1-i}{2}\right).$

Remark that if $p = 1$ then the Whishart reduces to a Gamma distribution

- suppose both μ and Σ^{-1} are unknown with prior distribution

$$\pi(\mu, \Sigma^{-1}) = \pi(\mu|\Sigma^{-1})\pi(\Sigma^{-1}).$$